Mohler and Porter *Crime Sci* (2021) 10:17
https://doi.org/10.1186/s40163-021-00152-x

Crime Science

SHORT CONTRIBUTION

Open Access

A note on the multiplicative fairness score in the NIJ recidivism forecasting challenge

George Mohler[1*] and Michael D. Porter[2]

Abstract

Background: The 2021 NIJ recidivism forecasting challenge asks participants to construct predictive models of recidivism while balancing false positive rates across groups of Black and white individuals through a multiplicative fairness score. We investigate the performance of several models for forecasting 1-year recidivism and optimizing the NIJ multiplicative fairness metric.

Methods: We consider standard linear and logistic regression, a penalized regression that optimizes a convex surrogate loss (that we show has an analytical solution), two post-processing techniques, linear regression with re-balanced data, a black-box general purpose optimizer applied directly to the NIJ metric and a gradient boosting machine learning approach.

Results: For the set of models investigated, we find that a simple heuristic of truncating scores at the decision threshold (thus predicting no recidivism across the data) yields as good or better NIJ fairness scores on held out data compared to other, more sophisticated approaches. We also find that when the cutoff is further away from the base rate of recidivism, as is the case in the competition where the base rate is 0.29 and the cutoff is 0.5, then simply optimizing the mean square error gives nearly optimal NIJ fairness metric solutions.

Conclusions: The multiplicative metric in the 2021 NIJ recidivism forecasting competition encourages solutions that simply optimize MSE and/or use truncation, therefore yielding trivial solutions that forecast no one will recidivate.

Introduction

The 2021 NIJ recidivism forecasting challenge is a competition hosted by the National Institute of Justice with the aim to "increase public safety and improve the fair administration of justice across the United States"[1]. The challenge focuses on data from the State of Georgia on individuals released from prison to parole supervision for the period January 1, 2013, through December 31, 2015. Challenge participants are tasked with constructing a predictive model of 1, 2, and 3 year recidivism upon release from prison based on variables such as age, gender, race, education, prior arrests and convictions, and other covariates.

The scoring metric used in a majority of categories of the competition is the mean square error (Brier score):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - p_i)^2, \quad (1)$$

where y_i is the binary recidivism outcome for individual i indicating recidivism ($y_i = 1$) or no recidivism ($y_i = 0$), p_i is the forecasted probability of recidivism, and N is the number of individuals in the dataset. Given recent concerns of bias of predictive models of recidivism, such as disparate false positive and negative rates across different racial/ethnic groups (Richard et al. 2018; Chouldechova 2017), the NIJ challenge includes a second set

*Correspondence: gmohler@iupui.edu
[1] Department of Computer and Information Science, Indiana University-Purdue University Indianapolis, Indianapolis, USA
Full list of author information is available at the end of the article

[1] https://nij.ojp.gov/funding/recidivism-forecasting-challenge.

of categories aimed at balancing low MSE while reducing the difference of false positive rates (denoted FP below) between groups of Black and white individuals in the data. In particular, contestants' models are scored according to the metric:

$$(1 - MSE)(1 - |FP_{Black} - FP_{white}|). \tag{2}$$

We refer to this metric as the "NIJFM" (NIJ Fairness Metric). False positive rates require a binary prediction defined by a cutoff, which in the NIJ competition is defined to be $p_i \geq 0.5$. Whereas the goal in the first set of categories is to minimize MSE, the goal in the second set of categories is to maximize the NIJFM (which occurs when the MSE and the difference of false positive rates are close to zero). Unlike a number of loss functions in the fairness-aware machine learning literature that have additive penalties to encourage some type of fairness (Yahav and Katrina 2017; Richard et al. 2017), the NIJFM in Eq. (2) is a multiplicative loss function defined by the product of the target loss (MSE) and the fairness penalty (difference in false positive rates). In this short note we explore several regression based methods for optimizing the NIJFM using data from the NIJ competition.

Methods

We consider several alternative linear regression models for optimizing the NIJFM. All linear models will be of the form,

$$p_i = X_i^t \theta, \tag{3}$$

where X_i^t is a covariate vector for individual i and θ is a vector of coefficients that we are optimizing. Thus differences in the models will be defined by differences in how the θ are estimated or by post-processing of the scores p_i.

The first approach is simply linear regression where the optimal coefficient vector for minimizing the MSE solves the linear equation:

$$\frac{2}{N} X^t X \theta - \frac{2}{N} X^t y = 0. \tag{4}$$

We also consider a balanced version of linear regression where the observations are re-weighted so that observations of each racial group contribute equally to the MSE. We also analyze linear regressions that are estimated on each racial group separately. We refer to these methods as linear reg., linear reg. (balanced) and linear reg. (group) respectively.

The next method, outlined in (Yahav and Katrina 2017; Richard et al. 2017), considers a convex surrogate loss where the step function representing the decision at the cutoff is replaced by a linear approximation (simply the score itself):

$$MSE + \lambda \left(\sum_{X_i \in S_{00}} \frac{X_i^t \theta}{|S_{00}|} - \sum_{X_i \in S_{10}} \frac{X_i^t \theta}{|S_{10}|} \right)^2. \tag{5}$$

Here S_{00} is the set of individuals of race 0 that did not recidivate ($y_i = 0$) and S_{10} is the set of individuals of race 1 that did not recidivate. The penalty term encourages the average scores over the negative class ($y_i = 0$) to be matched across race (as λ increases). This is a form of group fairness where we wish false positive rates to match across groups (alternatively individual fairness can be defined by bringing the summation outside of the squared term (Richard et al. 2017)). Because the loss function in Eq. (5) is quadratic, there is an analytical solution determined by the linear system:

$$\begin{aligned} &\left[\frac{2}{N} X^t X + 2\lambda (V_0^t V_0 - V_0^t V_1 - V_1^t V_0 + V_1^t V_1) \right] \\ &\theta - \frac{2}{N} X^t y = 0, \end{aligned} \tag{6}$$

where $V_j = \sum_{X_i \in S_{j0}} \frac{X_i^t}{|S_{j0}|}$. We select λ by choosing the value that yields the best NIJFM score on the training data. We refer to this method as the convex surrogate method.

Fairness can also be encouraged by post-processing the scores (Dennis et al. 2020). Here we use a simple shrinkage method where, for Black individuals[2] above the decision boundary cutoff (0.5 or greater for the NIJ competition), we subtract a constant value ϵ from their scores and then take the max of that value and the cutoff minus .0001.[3] We then choose the value of ϵ that optimizes the NIJFM on the training data. We refer to this method as linear regression with shrinkage. A second, even simpler, post-processing technique forces the false positive rates to zero by truncating all scores to the cutoff value (minus 0.0001) if they are above the cutoff. We refer to this method as linear regression with truncation.

While the step function representing the decision boundary in the NIJFM makes the metric non-continuous and non-differentiable, nonetheless one can attempt to optimize it with general purpose optimization software. We find that the optim function in the R stats library works reasonably well at optimizing the NIJFM when given the linear regression coefficients as an initial guess (and using the "BFGS" method, a quasi-Newton method with finite difference approximations for derivatives). We refer to this method as BFGS.

[2] Shrinkage is applied to the group with higher false positive rates, which in the NIJ competition corresponds to Black individuals.

[3] Under the competition rules, scores are rounded to 4 decimal places and scores of 0.5000 and above are considered the positive class. Therefore we shrink scores down to 0.4999.

Table 1 Mean square error (MSE), false positive rates (FP) by race, and NIJFM scores on held-out (50%) test data with competition cutoff of 0.5 for decision boundary. Bootstrap standard errors reported in parentheses

Model	MSE	FP (black)	FP (white)	NIJFM
Linear Reg.	0.192 (0.002)	0.048 (0.004)	0.030 (0.003)	0.793 (0.005)
Logistic Reg.	0.192 (0.002)	0.067 (0.004)	0.042 (0.004)	0.787 (0.005)
Linear Reg. (Trunc.)	0.192 (0.002)	0	0	0.808 (0.002)
Linear Reg. (Shrink)	0.192 (0.002)	0.034 (0.003)	0.030 (0.003)	0.804 (0.004)
Convex Surrogate	0.193 (0.002)	0.043 (0.003)	0.026 (0.003)	0.794 (0.004)
BFGS	0.193 (0.002)	0.035 (0.003)	0.021 (0.003)	0.796 (0.004)
Linear Reg. (Balanced)	0.192 (0.002)	0.048 (0.004)	0.030 (0.003)	0.793 (0.005)
Linear Reg. (Group)	0.192 (0.002)	0.045 (0.004)	0.033 (0.003)	0.798 (0.005)
XG Boost	0.192 (0.002)	0.045 (0.004)	0.030 (0.003)	0.796 (0.005)
XG Boost (Trunc.)	0.193 (0.002)	0	0	0.807 (0.002)
XG Boost (Shrink)	0.192 (0.002)	0.033 (0.003)	0.030 (0.003)	0.804 (0.003)

Finally, we implemented a logistic regression to investigate the effects of using a binomial likelihood instead of Gaussian and a gradient boosting model (xgboost (Tianqi and Carlos 2016)) to explore the performance of a nonlinear machine learning approach. Hyper-parameters for the convex surrogate, shrinkage, and xgboost models are tuned using a grid search on the training data and model performance is evaluated on held-out test data. The code to reproduce the results is available on github.[4]

Data

The data used in this study comes from the NIJ recidivism forecasting challenge website[5] and comprises 18,028 individuals released from prison to parole supervision in the state of Georgia for the period January 1, 2013, through December 31, 2015. We split the data provided by the competition into a training set that we use to build our models (consisting of 9000 randomly sampled rows) and use the remaining rows as a hold-out test set. The overall data consists of 10,313 Black individuals and 7715 white individuals. Of the 18,028 total individuals, 5377 individuals are labeled as having recidivated in year 1 following release (hence the base rate is $5377/18{,}028 \approx 0.298$). We note that the competition also focuses on year 2 and 3, along with separate scores for women and men, however we restrict our attention to overall MSE and NIJFM (aggregated across men and women) and focus on year 1 recidivism. The covariates we use to construct our models include: gender (sex), race (Black or white), an indicator for being gang affiliated, age at release from prison, years spent in prison, total prior arrests, total prior convictions, education level, and number of dependents.

Results

In Table 1 we display results for the models from Sect. 2 on held-out test data for the NIJ competition using a decision threshold of 0.5. We include the MSE, false postive rates (FP), and the NIJFM scores along with bootstrap standard errors.[6] Differences in the MSE across models are not statistically significant, with all models achieving a held-out MSE of 0.192–0.193. NIJFM scores range from 0.787 to 0.808, with simple truncation or shrinkage applied to linear regression and xgboost having as good or better fairness scores compared to the other approaches. While the models with truncation or shrinkage yield good NIJFM scores, we note that these models violate individual fairness since they involve post-processing of only a subset of individuals' scores. Also, regression and boosting with truncation yield a trivial solution with no individuals predicted to recidivate (though such a solution may very well win the fairness category under the competition design).

We note that the false positive rates in Table 1 are low across all methods. This is due to the fact that the decision cutoff of 0.5 is far from the base rate of recidivism for the dataset (0.298). To investigate the sensitivity of results to the decision cutoff further, in Table 2 we display results for the models from Sect. 2 on held-out test data for cutoffs of 0.3 (corresponding to more false positives and less false negatives) and 0.7 (corresponding to less false positives and more false negatives). As the cutoff moves further away from the base rate of recidivism, we see less of a difference in NIJFM scores across fairness-aware regressions and the standard linear/logistic regressions. This is because fewer individuals are forecasted to

[4] https://github.com/gomohler/NIJ-Recidivism-Forecasting.

[5] https://nij.ojp.gov/funding/recidivism-forecasting-challenge.

[6] Bootstrap standard errors are calculated for each model fit to the training data by resampling the test data with replacement 1000 times and calculating the standard deviation of the statistic across samples.

Table 2 NIJFM scores and false positive rates on held-out (50%) test data with cutoffs of $c = 0.3$ and $c = 0.7$ for decision boundary. Bootstrap standard errors reported in parentheses

Model	FP black ($c = 0.3$)	FP white ($c = 0.3$)	NIJFM ($c = 0.3$)	FP Black ($c = 0.7$)	FP white ($c = 0.7$)	NIJFM ($c = 0.7$)
Linear Reg.	0.485 (0.009)	0.354 (0.009)	0.702 (0.010)	0	0	0.808 (0.002)
Logistic Reg.	0.430 (0.008)	0.303 (0.009)	0.706 (0.010)	0.003 (0.001)	0.001 (0.001)	0.807 (0.002)
Linear Reg. (Trunc.)	0	0	0.798 (0.002)	0	0	0.808 (0.002)
Linear Reg. (Shrink)	0.369 (0.008)	0.354 (0.009)	0.795 (0.008)	0	0	0.808 (0.002)
Convex Surrogate	0.478 (0.009)	0.377 (0.009)	0.725 (0.010)	0	0	0.808 (0.002)
BFGS	0.434 (0.008)	0.428 (0.010)	0.795 (0.007)	0	0	0.808 (0.002)
Linear Reg. (Balanced)	0.487 (0.008)	0.354 (0.009)	0.701 (0.010)	0	0	0.808 (0.002)
Linear Reg. (Group)	0.485 (0.008)	0.358 (0.009)	0.705 (0.010)	0	0	0.808 (0.002)
XG Boost	0.472 (0.008)	0.377 (0.009)	0.731 (0.010)	0	0	0.808 (0.002)
XG Boost (Trunc.)	0	0	0.798 (0.002)	0	0	0.808 (0.002)
XG Boost (Shrink)	0.386 (0.008)	0.377 (0.009)	0.798 (0.007)	0	0	0.808 (0.002)

be above the decision boundary, and therefore the false positive rates are much lower (and approach zero as the decision boundary moves further from the base rate).

Conclusion

We fit several linear and gradient boosting models to data from the NIJ recidivism forecasting challenge that were optimized in different ways to match false positive rates across Black and white groups of individuals. We found that a simple heuristic of cutting scores at a value slightly below the decision threshold (thus predicting no recidivism across the data) yields as good or better NIJFM scores on held out data compared to the other, more sophisticated approaches that we considered. Such a solution is not useful in practice; practical scores need to include values above the decision threshold in order to identify high-risk individuals in need of greater support resources or other interventions. We also found that when the decision cutoff for recidivism is further away from the base rate of recidivism, as is the case in the competition, then simply optimizing the MSE also gives a nearly optimal NIJFM score.

Our analysis has several limitations. First, we restricted our attention to linear models and gradient boosting on a subset of competition features. Other types of statistical or machine learning approaches applied to an expanded feature set may yield different results. Second, we restricted our attention to year 1 recidivism on data aggregated across sex/gender.

Finally, we note that NIJ released results for the competition on July 16, 2021.[7] Our boosting-ensemble based submission, under the team name PASDA, used truncation for the female category and no truncation for the male category. Our submission placed 1st, 3rd and 2nd in the three fairness category rounds for female recidivism, while not placing in the top 5 in any round for the fairness category of male recidivism. This provides additional support for the conclusions drawn in the present paper.

On the MSE as a loss function for training models

We note that it is more common to use a cross-entropy loss for training binary classifiers, rather than using MSE. For a good treatment of this issue, see (Zhang 2004) where Zhang shows that regression with a mean square error loss (and then truncating at 0/1) is consistent. In the present paper, we observe similar performance in terms of the MSE and NIJFM scores between linear and logistic regression.

On alternative metrics for evaluation of recidivism forecasts

While the MSE can be used to estimate recidivism forecasting models, other metrics may be better suited for model evaluation. In practice, confusion tables that contain accuracy, false positive rates, and false negative rates will better highlight the tradeoffs between different models and cutoff choices. Given that results can change depending on the choice of cutoff, comparing cost, ROC, and precision-recall curves (see Fig. 1) may provide more insight than a single metric. Also, the type and severity of the crime committed could be incorporated into a metric, similar to how the gini index is used in evaluating insurance risk models (Frees et al. 2011). Finally, we note that there are alternative definitions of fairness that have been discussed in the literature (Ninareh et al. 2019; Sam and Sharad 2018), and these may provide a more

[7] https://nij.ojp.gov/funding/recidivism-forecasting-challenge-results.

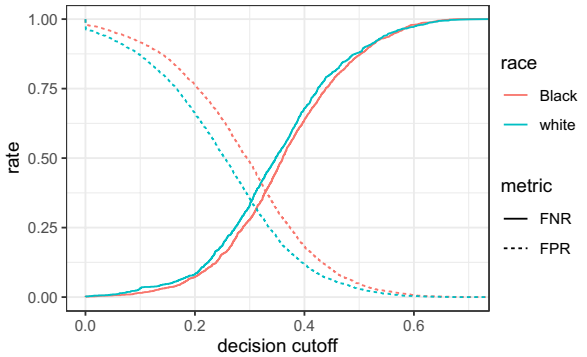

Fig. 1 False positive and negative rate vs. decision cutoff by race for linear regression

nuanced assessment of recidivism forecasts than group false positive rates.

Acknowledgements
We thank the reviewers for their insightful comments and suggestions. This research was supported in part by NSF grant SCC-1737585 and the IU Racial Justice Research Fund. GM is a director on the board of Geolitica, an analytics company serving law enforcement.

Author details
[1]Department of Computer and Information Science, Indiana University-Purdue University Indianapolis, Indianapolis, USA. [2]School of Data Science and Department of Engineering Systems and Environment, University of Virginia, Charlottesville, USA.

Received: 18 June 2021 Accepted: 28 July 2021
Published online: 06 August 2021

References

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163.
Dennis, W., Karthikeyan, N.R., & Flavio, C. (2020). Optimized score transformation for fair classification. In *International Conference on Artificial Intelligence and Statistics*, 1673–1683. PMLR
Frees, E. W., Meyers, G., David, A., & Cummings. (2011). Summarizing insurance scores using a Gini index. *Journal of the American Statistical Association, 106*(495), 1085–1098.
Ninareh, M., Fred, M., Nripsuta, S., Kristina, L. & Aram, G. (2019). A survey on bias and fairness in machine learning. *arXiv preprint* arXiv:1908.09635.
Richard, B., Hoda H., Shahin, J., Matthew, J., Michael, K., Jamie, M., Seth, N. & Aaron, R. (2017). A convex framework for fair regression. In: *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning. (FATML) 2017*.
Richard, B., Hoda, H., Shahin, J., Michael, K. & Aaron, R. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 0049124118782533.
Sam, C.-D., & Sharad, G. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint*. arXiv:1808.00023.
Tianqi, C. & Carlos, G. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, Association for Computing Machinery.
Yahav, B. & Katrina, L. (2017). Penalizing unfairness in binary classification. *arXiv preprint*. arXiv:1707.00044.
Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics, 32*(1), 56–85.

Publisher's Note